



가장 쉽고 안전한 엔터프라이즈 AI의 시작 : VCF 9.1 기반 Private AI Service

이청수 부장

브로드컴 (VMware by Broadcom)



Why Private AI & Why VCF?

AI는 '실험'에서 '운영'으로의 전환 시기



88%

1개 이상 업무에 AI 정기 사
용
(전년 78%)



약 1/3

전사 확장 단계 도달
— 2/3는 여전히 파일럿



51%

AI 사용 중 1회 이상
부정적 결과 경험



3.1개

기업당 평균 모델 프로바이더
— 비용·통제가 동인

핵심

AI 도입은 보편화(88%) 다만 약 30%만 전사 확장, 진짜 과제는 '파일럿→운영' 전환,
토큰 비용 위험 도래

출처: McKinsey, The State of AI 2025 (2025.11, n=1,993) / ICONIQ, 2026 State of AI Bi-Annual Snapshot (n≈300)



Why Private AI & Why VCF?

프롬프트와 추론 자체가 데이터 유출 경로



KPMG

69%

AI 데이터 프라이버시를
우려하는 임원 비율
(KPMG 보고서)



프라이버시 = AI 윤리 우려 1위

전문가 ~40%가 1순위로 지목,
2023년 대비 약 2배로 증가함 (Deloitte)

Deloitte.



민감정보가 GenAI로 새어 나간다

프롬프트 4.37%, 업로드 파일 22%에서 탐지됨
(소스코드, 자격증명, 고객 기록 포함)

Cyber
Security
Research



유출의 40%+가 AI에서 비롯

2027년까지 기업 데이터 유출의 40% 이상이 부적절하거나
미승인된 GenAI 사용으로 발생할 것으로 예측됨
(Gartner)

Gartner

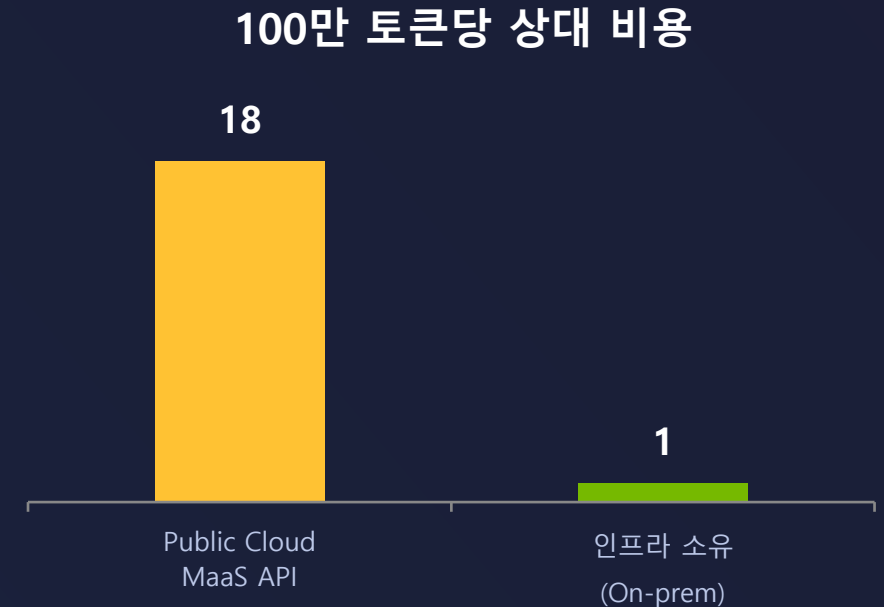
출처: AI Quarterly, Pulse Survey: Q2 2025, June 26, 2025(KPMG), State of Ethics and Trust in Technology, 2024(Deloitte),
The State of Browser Security Report 2026 (Keep Aware), Gartner Predicts 40% of AI Data Breaches Will Arise from Cross-Border GenAI Misuse by 2027(Gartner)



Why Private AI & Why VCF?

'무소유 모델'은 운영 단계에서부터 부채

- ▶ 토큰 과금(MaaS)은 사용량 증가에 따라 비용이 선형으로 폭증
- ▶ Egress(데이터 송신) 비용이 클라우드 AI 지출의 15~30%
→ 프라이빗 클라우드 환경은 = 0
- ▶ 인프라를 '소유'하면 100만 토큰당 최대 18배 비용 우위*
- ▶ 고활용 추론에서는 4개월 내 손익분기 도달



핵심

LLM API 지출은 1년 새 2배(\$3.5B→\$8.4B), 엔터프라이즈 추론 청구서는 예산의 2~4배로 폭증

출처: *Lenovo Press, On-Premise vs Cloud GenAI TCO 2025/2026 / aisuperior-Codezilla LLM Cost in Production 2026





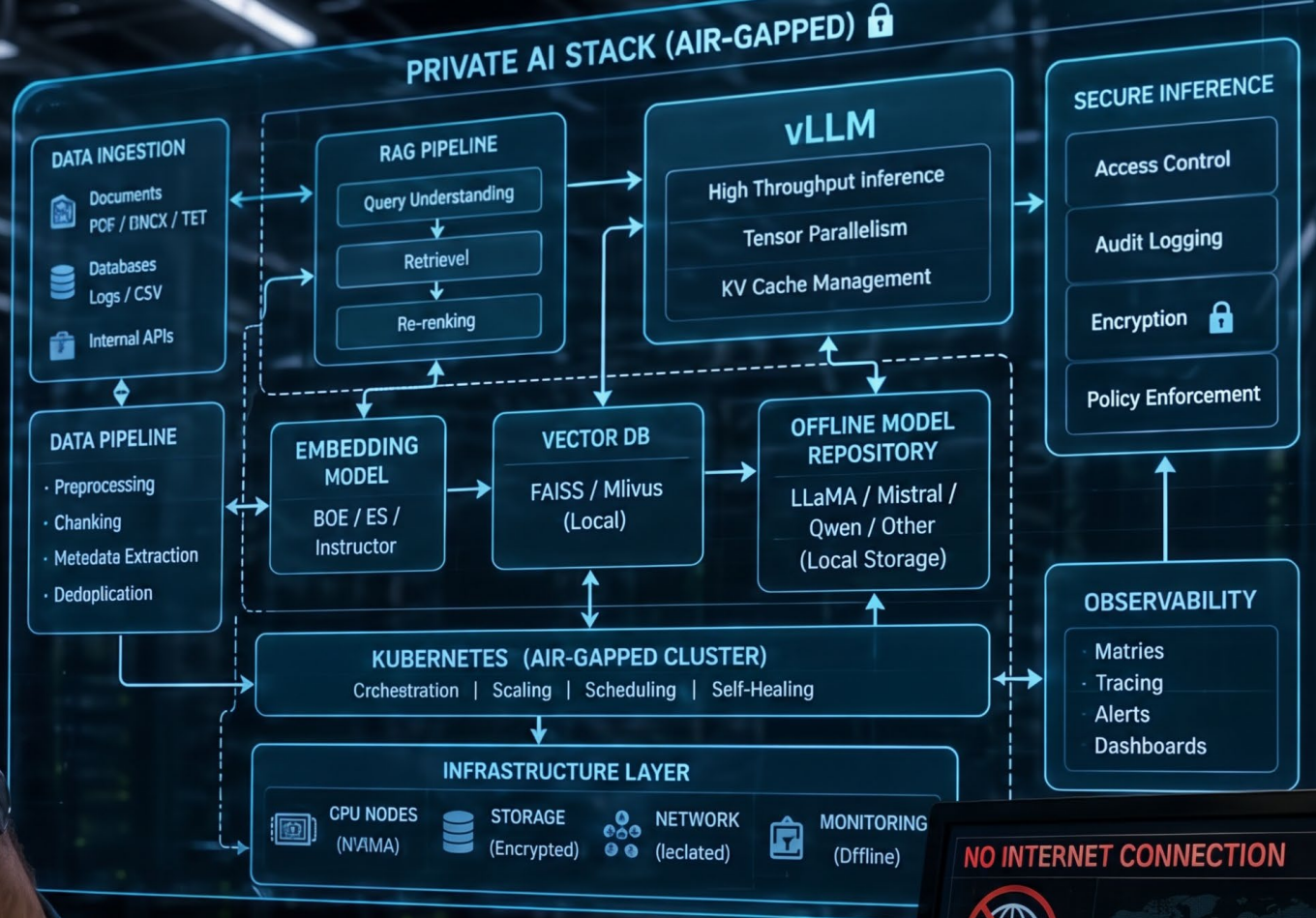
AIR-GAPPED AI ENVIRONMENT

- NO INTERNET ACCESS
- NO EXTERNAL NETWORKS
- ALL DATA STAYS INSIDE
- MAXIMUM SECURITY

```
INFO Starting vLLM engine (tomot_oeralla1_alicena)
INFO Leading model sherde...
INFO Allocating UD cache...
WARN CUDA memory usage high
ERROR NCCL error: unhandled pyteam error
ERROR Failed to initialize NCCL communicator
INFO Retrying in 30 seconds...
ERROR GDM: Out of memory while allocating tessccr
ERROR vLLM ampine crashed
INFO Restarting inference engine...
INFO Loading model shards...
ERROR Vector DB connection timeout
ERROR Failed to connect to local hilooos intotence
WARN High latency in retrieval pipnlco
ERROR RAB pipeline failed: oo documents rooioeust
ERROR Embedding model lead failed (bceetons simmetot)
ERROR All martr failed to load
ERROR Inference request failed
ERROR Service onavailable
```

REMEMBER:
WHY YOU'RE
THIS

BUILDING



NO INTERNET CONNECTION



AIR-GAPPED MODE

All operations are executed in isolated mode

NETWORK STATUS

DISCONNECTED

EXTERNAL ACCESS

LOCKED



CHALLENGES

- Model Compatibility
- GPU Memory Limits
- KV Cache Management
- Tensor Parallelism
- NCCL Errors

SIMPLIFY. ACCELERATE. TRANSFORM.

VMware-By-Broadcom & NVIDIA: The Unified Platform Built for AI.

Unified VMware-By-Broadcom Infrastructure

1. Cloud-Native Compute
2. Scalable Storage
3. Broadcom-Powered Networking
4. Zero-Trust Security
5. End-to-End Management

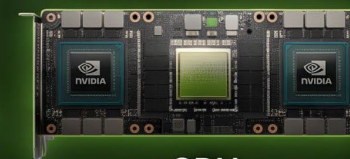
Private Cloud

&
VMware Cloud Foundation
NVIDIA AI Enterprise

Powered by
Broadcom  NVIDIA.

NVIDIA AI Software & Accelerated Computing

1. AI Frameworks
2. Pre-trained Models
3. AI Enterprise Tools
4. Multi-GPU Optimization
5. Global Support



GPU
Broadcom I/O Optimized



One-Click
Deployment



Single-Pane
Management



Predictable
Scaling

Integrated Technology. Unified Experience. Infinite Possibilities.

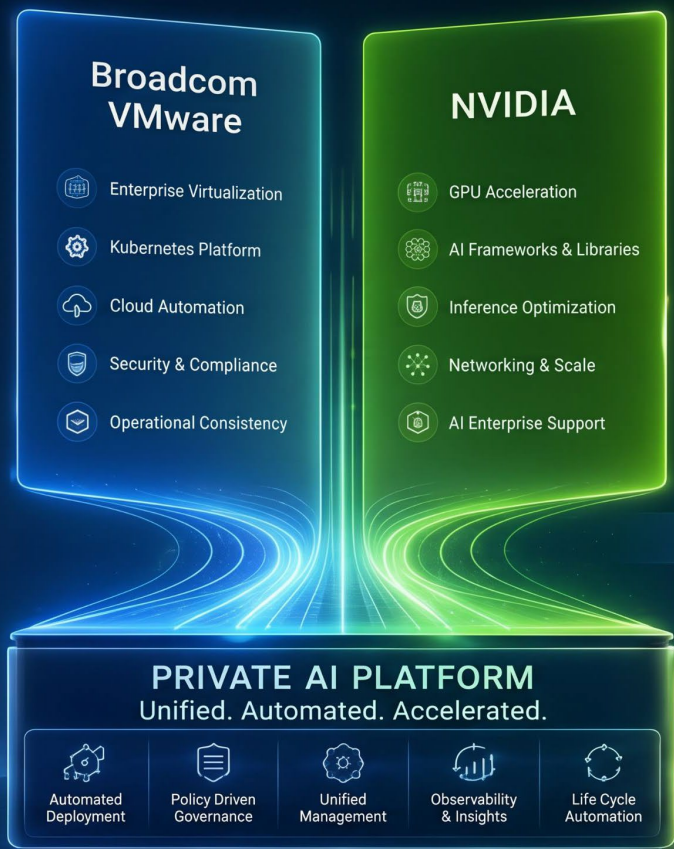
BEFORE

Complex. Siloed. Hard to Operate.



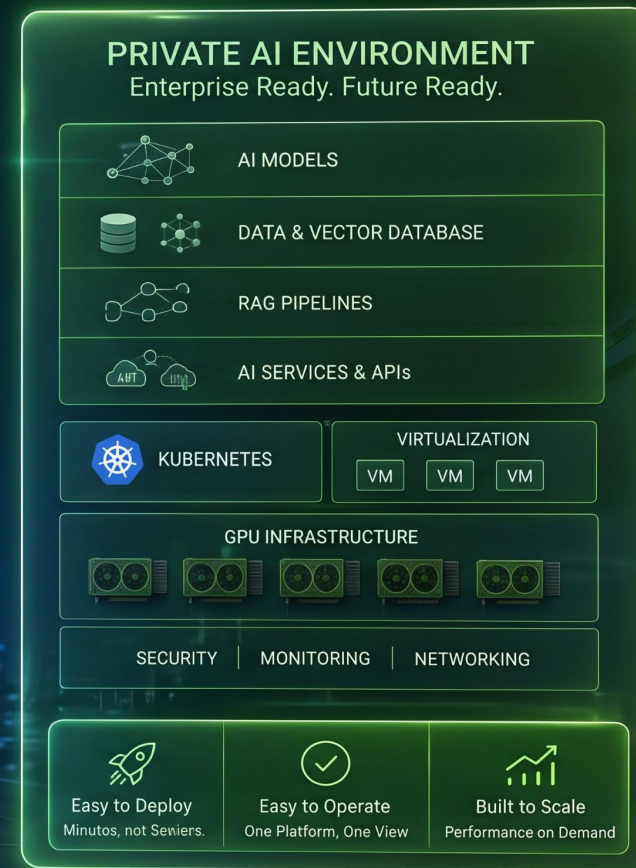
THE UNIFICATION LAYER

One Platform. Powered by Two Leaders.



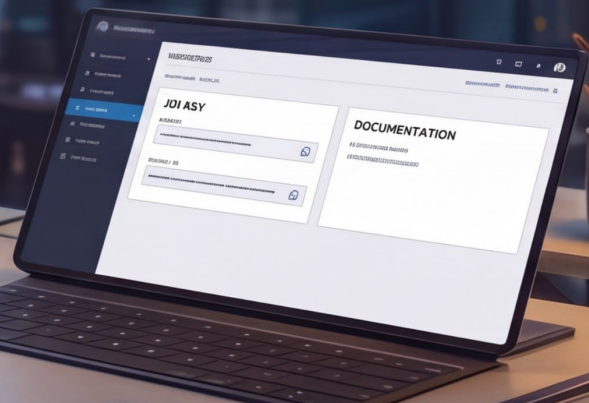
AFTER

Simple. Unified. Easy to Operate.



ACCELERATE AI. SIMPLIFY OPERATIONS. DELIVER ENTERPRISE VALUE.





VCF Private AI Services

AI 인프라에서부터 서비스까지, 단일 플랫폼에서 완성



Enable
Privacy & Security



Simplify
Infrastructure Setup



Streamline
Model Deployment



Unlock gen AI & Unleash Productivity, With Lower TCO

사내 데이터만 믿는다 – 실질적인 RAG 챗봇 구현

일반적인 지식이 아닌, '우리 회사 매뉴얼'을 읽는 AI



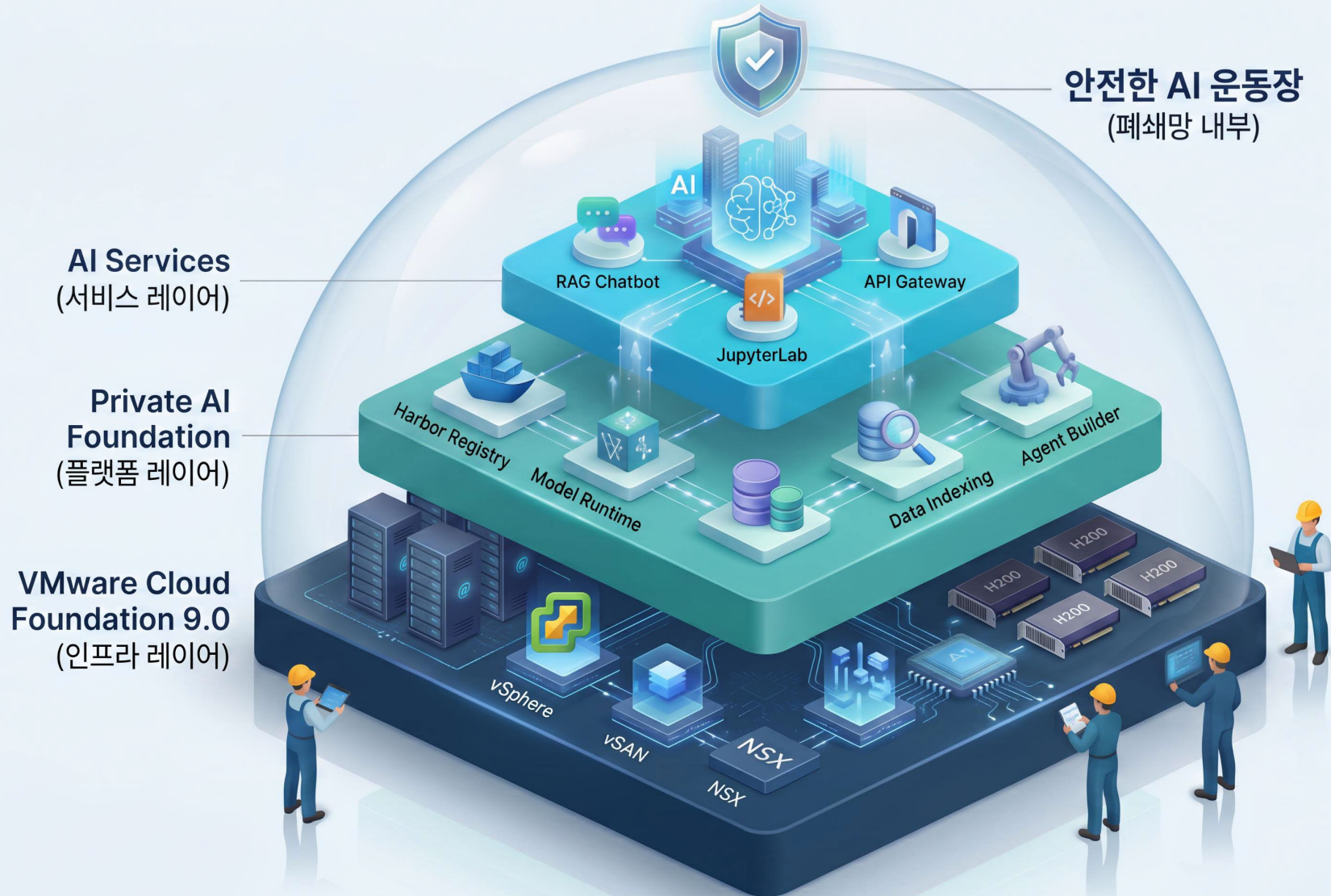
✓ 폐쇄망 Harbor 구축 완료

✓ 사내 문서 KB화

✓ RAG 신뢰도 확보

VCF + PAIF – 단단한 토대 위에 세운 AI

VMware Cloud Foundation 9+ Private AI Foundation 통합 아키텍처



안전한 AI 운동장
(폐쇄망 내부)

AI Services
(서비스 레이어)

Private AI
Foundation
(플랫폼 레이어)

VMware Cloud
Foundation 9.0
(인프라 레이어)



GPU 리소스 효율 관리

가상화 기술로 GPU 자원을
최적화하여 효율적으로 관리



폐쇄망 내 일관된 운영 환경

Air-Gap 환경에서 일관된
AI 개발 및 운영 환경 제공



인프라-플랫폼 통합

VCF 9.0과 PAIF의 결합으로
완벽한 인프라-플랫폼 통합 실현

VMware Private AI Service 기능 검증 고객사 기능 검증 수행

배경 및 목적

- ✓ 현 운영 서버의 vSphere 8.x 지원 종료
- ✓ VCF 9.0 기반 Private AI Foundation 실환경 검증
- ✓ 데이터 AI 분석 환경 구축 요구
- ✓ 임대 장비 활용 PoC

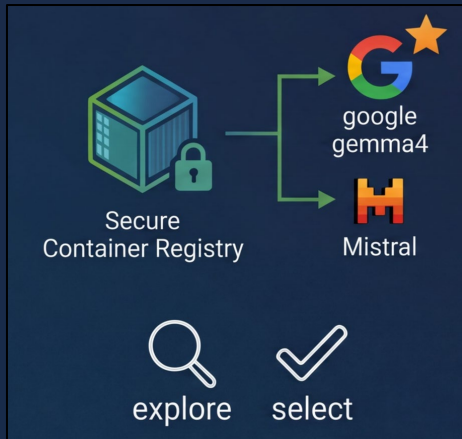
PoC 검증 범위

- ✓ VCF 9.0 MGMT Domain + WL Domain 구성
- ✓ Private AI Foundation 설치 및 기능 구성
- ✓ GPU 지원 VKS Cluster 자동 프로비저닝
- ✓ NSX를 통한 네트워크 구성
- ✓ AI 데모 시연



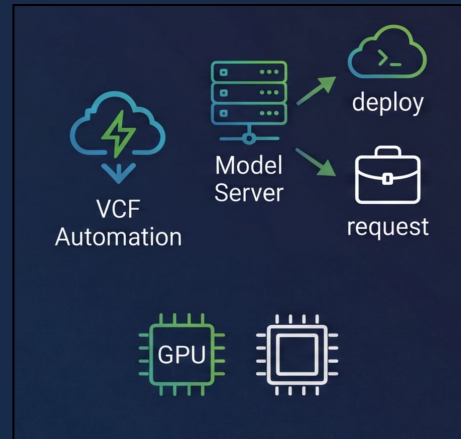
4 Simple Step

1. 모델 탐색 및 선택



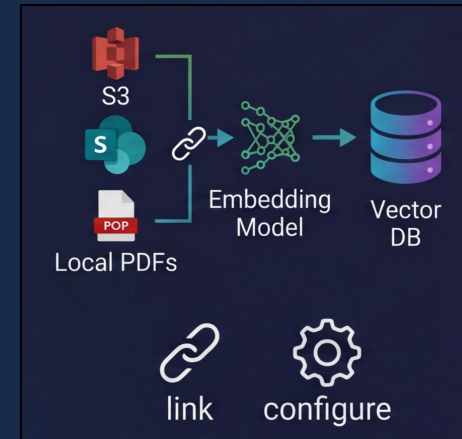
필요한 LLM/임베딩 모델을 Harbor 레지스트리에서 탐색 LLM 모델별 버전/형상 관리

2. 모델 런타임 배포



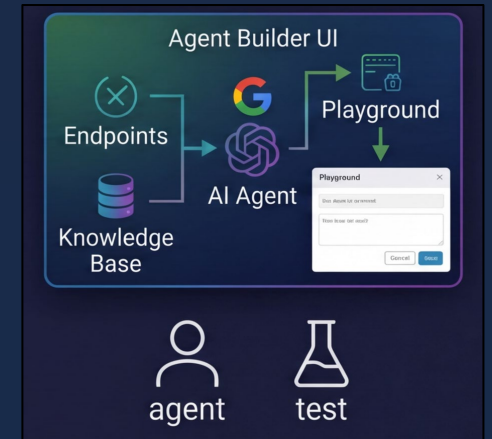
VCF Automation 포털에서 선택한 모델의 전용 런타임 배포 요청

3. 지식 기반(KB) 구성



Private AI 서비스 포털에서 지식 기반(Knowledge Base) 구성 및 데이터 연동 설정

4. AI Agent 배포/테스트/운영관리



Agent Builder UI 기반으로 배포된 모델과 KB를 결합하여 맞춤형 Agent 생성, Playground로 구성 테스트



VMware Private AI Service 기능 #1

모델 스토어 (Harbor)

- Llama-3.1-70b/Mixtral 8x22b 등 대/소규모 언어 모델 활용
- 외부 유출 없이 사내 인프라 내에서 모델 형상관리 수행
- vLLM 모델 런타임을 통한 서비스화
- Private AI Foundation 기반 모델 서빙

KEY POINT

- 오픈소스 모델 활용 MLOps/AIOps 구현
- VKS 기반 자동 배포 및 오케스트레이션, 챗봇 서비스

The screenshot shows the Harbor web interface for the 'models' project. The left sidebar contains navigation links: Projects, Logs, Administration (Users, Robot Accounts, Registries, Replications, Distributions, Labels, Project Quotas, Interrogation Services, Clean Up, Job Service Dashboard, Configuration), DARK, and Harbor API V2.0. The main content area shows the 'models' project details, including 'Access Level: Public' and 'Quota used: 226.40GiB of unlimited'. Below this is a table of repositories with columns for Name, Artifacts, Pulls, and Last Modified Time.

Name	Artifacts	Pulls	Last Modified Time
models/mazyarpanahi/mixtral-8x22b-instruct-v0.1-awq	1	2	3/21/26, 2:28 PM
models/mistralai/mistral-7b-instruct-v0.3	1	6	4/3/26, 2:17 PM
models/dragonkue/bge-m3-ko	1	20	4/2/26, 5:00 PM
models/upskeyy/bge-m3-korean	1	5	4/3/26, 3:14 PM
models/hugging-quants/meta-llama-3.1-70b-instruct-awq-int4	1	6	4/3/26, 10:38 AM
models/mistralai/mistral-small-3.1-24b-instruct-2503	1	2	3/21/26, 4:34 AM



VMware Private AI Service 기능 #1

LLM Model 배포

- VCF Automation에서 vGPU 쿠버네티스 환경(VKS) 자동 프로비저닝 → 원클릭 클러스터 배포 자동화 구현
- Nvidia H200 기반 GPU 프로파일 선택 → 최적의 컴퓨팅 자원 할당

The screenshot displays the VMware Cloud Foundation Automation interface during the deployment of an LLM model. The left sidebar shows the navigation menu with 'Private AI' selected. The main panel is divided into three sections: 'Model Information', 'Endpoint Specification', and 'Deployment Configuration'. The 'Deployment Configuration' section is active, showing a table of VM Class options. The 'GPU-Enabled' radio button is selected. The table lists five VM Class options, with 'h200-1-18g-c12-m36' selected. The 'Storage Class' is set to 'vsan-wld-policy01'. A 'NEXT' button is visible at the bottom of the configuration section. On the right, the 'Kubernetes Resource YAML' section shows the generated YAML for the 'Model Endpoint' resource, including details like 'apiVersion', 'kind', 'spec', and 'metadata'.

VM Class	CPU	Memory	vGPU Profile
☒ h200-1-18g-c12-m36	12	36 GiB	nvidia_h200-1-18c
☐ h200-2-35g-c32-m70	32	70 GiB	nvidia_h200-2-35c
☐ h200-3-71g-c32-m141	32	141 GiB	nvidia_h200-3-71c
☐ h200-4-71g-c32-m160	32	160 GiB	nvidia_h200-4-71c

```
1 apiVersion: pais.vmware.com/v1alpha1
2 kind: ModelEndpoint
3 spec:
4   model:
5     ociRef: harbor.vcfdemo.vm/models/upskey
6     pullSecrets:
7       - name: pais-pais-reg-creds-6b40fed3
8     routingName: upskey/bge-m3-korean
9     type: Embeddings
10    engine: Infinity
11    virtualMachineClassName: h200-1-18g-c12-m36
12    storageClassName: vsan-wld-policy01
13 metadata:
14   name: upskey
15   namespace: k
16   annotations:
17     pais.vmware.com/displayName: upskey/bge-m3-korean
18
```



VMware Private AI Service 기능 #1

LLM Model 배포

- 배포간 선택된 LLM/임베딩 모델 엔진, 인프라 사양, vSAN 스토리지 정책 확인
- 배포 프로세스의 일관성 및 자동화 구현

The screenshot displays the VMware Cloud Foundation Automation interface. The top navigation bar includes 'Overview', 'Build & Deploy', 'Manage & Govern', 'Orchestrate', 'Administer', and 'Inbox'. The left sidebar shows a tree view with 'Instances', 'Catalog', 'Content Hub', 'Blueprint Design', 'VM Images', 'Content Libraries', 'Workflows', 'Event Subscriptions', 'Services', and 'Private AI'. The main panel is titled '4. Review and Deploy' and contains the following information:

Model Information	
Model URL	harbor.vcfdemo.vm/models/upskeyy/bge-m3-korean
Image Tag or Digest	v1
Description	

Endpoint Specification	
Routing Name	upskeyy/bge-m3-korean
Display Name	upskeyy/bge-m3-korean
Model Type	Embeddings
Engine	Infinity
Arguments for Engine Customization	None
Environment Variable Overrides	None
Override Configuration	None

Deployment Configuration	
VM Class Type	GPU
VM Class	h200-1-18g-c12-m36
Storage Class	vsan-wld-policy01

A green button labeled 'DEPLOY MODEL ENDPOINT' is located at the bottom of the deployment configuration section.

On the right side, the 'Kubernetes Resource YAML' panel shows the following configuration:

```
1 apiVersion: pais.vmware.com/v1alpha1
2 kind: ModelEndpoint
3 spec:
4   model:
5     ociRef: harbor.vcfdemo.vm/models/upskeyy/bge-m3-korean
6     pullSecrets:
7       - name: pais-pais-reg-creds-6b40fed3
8     routingName: upskeyy/bge-m3-korean
9     type: Embeddings
10    engine: Infinity
11    virtualMachineClassName: h200-1-18g-c12-m36
12    storageClassName: vsan-wld-policy01
13  metadata:
14    name: upskeyy-bge-m3-korean
15    namespace: knoc-ns-ybj6r
16    annotations:
17      pais.vmware.com/displayName: upskeyy/bge-m3-korean
```



VMware Private AI Service 기능 #1

LLM Model 배포

- 통합 대시보드 (GUI) 및 CLI 기반 실시간 서비스 가용성 확인
- 배포된 모델의 GPU VM 할당 상태와 쿠버네티스 리소스 생성 결과 최종 검증

The screenshot displays the VMware Cloud Foundation Automation console. The top navigation bar includes 'Overview', 'Build & Deploy', 'Manage & Govern', 'Orchestrate', 'Administer', and 'Inbox'. The left sidebar shows 'Instances', 'Catalog', 'Content Hub', and 'Services'. The main area is titled 'Private AI Services' with sub-tabs for 'Model Runtime' and 'Service Configuration'. A 'NEW MODEL ENDPOINT' button is visible. Below it is a table listing model endpoints:

Display Name	Status	Type	VM Class	VM Class Type
dragonkue/bge-m3-ko	Available	Embeddings	h200-1-18g-c12-m36	GPU
hugging-quants/meta-llama-3.1-70b-instruct-awq-int4	Available	Completions	h200-3-71g-c32-m141	GPU
upskyy/bge-m3-korean	Pending	Embeddings	cpu8-mem16	NonGPU

Below the table is a 'Manage Columns' button and a pagination indicator showing '3 objects'.

At the bottom, a terminal window shows the output of the command `kubectl get paisconfiguration -A; echo ""; kubectl get modelendpoints -A`. The output is as follows:

```
Every 3.0s: kubectl get paisconfiguration -A; echo ""; kubectl get modelendpoints -A
knoc-harbor.cloud.ds: Fri Apr 3 15:10:00 2026 ^

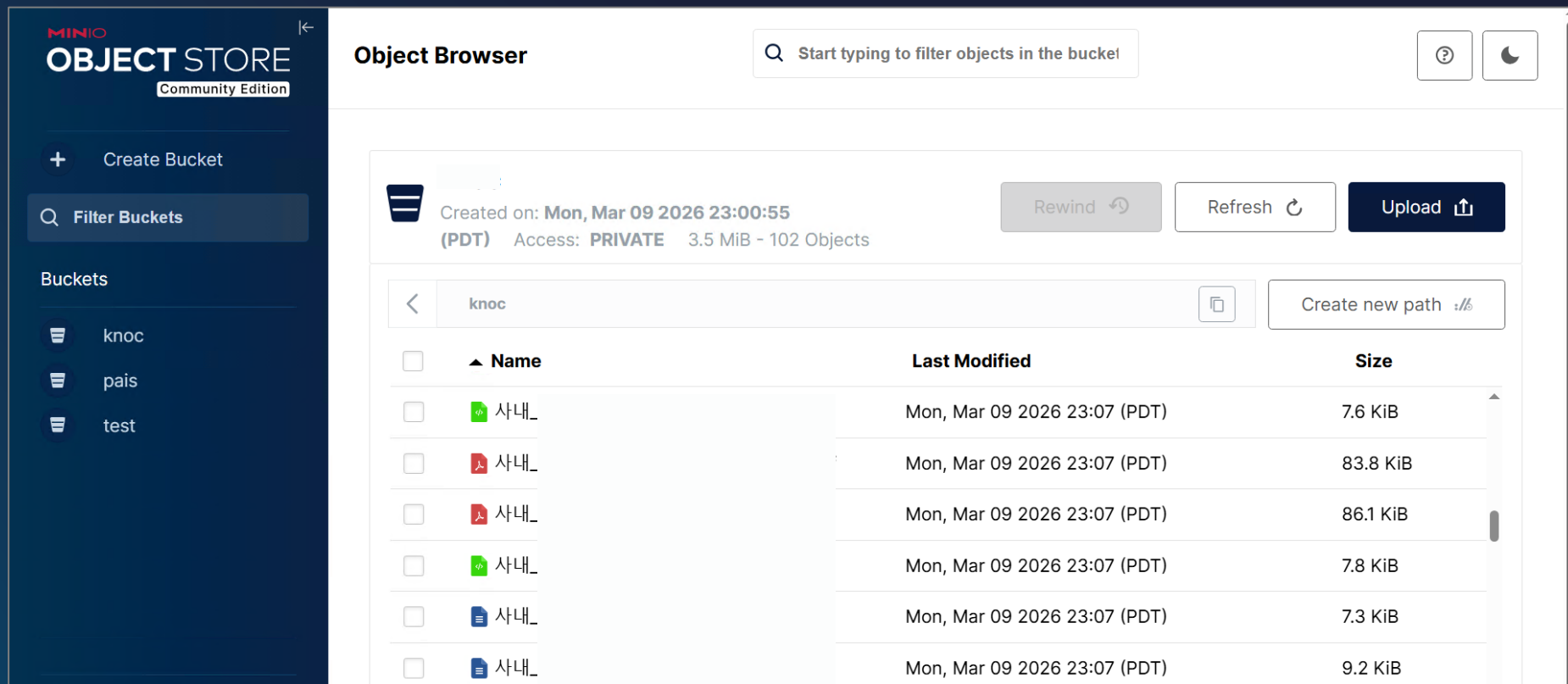
NAMESPACE NAME SERVICE READY REASON
knoc-ns-ybj6r default pais-ingress-default True paisAvailable

NAMESPACE NAME REPLICAS READY REASON MODEL
knoc-ns-ybj6r dragonkue-bge-m3-ko 1 True reconcileSucceeded harbor.vcfdemo.vm/models/dragonkue/bge-m3-ko:v1
knoc-ns-ybj6r hugging-quants-meta-llama-3.1-70b-instruct-awq-int4 1 True reconcileSucceeded harbor.vcfdemo.vm/models/hugging-quants/meta-llama-3.1-70b-instruct-awq-int4:v1
knoc-ns-ybj6r upskyy-bge-m3-korean 1 False waitingForVKSMachineDeployment harbor.vcfdemo.vm/models/upskyy/bge-m3-korean:v1
```



도메인 전문 지식 베이스 (Knowledge Base) 구성

- 데이터소스 준비 – 다양한 형식의 문서를 MinIO 버킷에 업로드하여 학습 데이터셋 구축
- 파일 포맷 : PDF, HTML, MD ,PPTX, DOCX, TXT, CSV



VMware Private AI Service 기능 #1

도메인 전문 지식 베이스 (Knowledge Base) 구성

- 데이터소스 연결 – 외부 데이터소스를 VMware Private AI Service에 연동

The screenshot shows the VMware Private AI Services interface. The left sidebar contains navigation options: Data Indexing and Retrieval, Knowledge Bases, Data Sources, MCP, MCP Servers, Tools Gallery, Agent Builder, Agents, Observability, and Metrics. The main content area is titled 'Data Sources' and includes a table of existing data sources and a form to add a new one.

Name	Source type	Last updated	Related knowledge bases
: Sample Data	Amazon S3 Compatible Store	Apr 2, 2026, 4:59:34 PM	4
krpdfest	Amazon S3 Compatible Store	Apr 3, 2026, 2:56:07 PM	1
pais test	Amazon S3 Compatible Store	Apr 3, 2026, 3:06:01 PM	1
krtest	Amazon S3 Compatible Store	Apr 3, 2026, 3:29:13 PM	1
engtest	Amazon S3 Compatible Store	Apr 3, 2026, 3:34:42 PM	1

Below the table is a 'Manage Columns' button and a pagination control showing 'Objects per page 10' and '5 objects'.

The 'Add Data Source' form is titled '2. Data source details' and includes the following fields:

- Data source type**: Amazon S3 Compatible Store
- URL**: https://minio.vcfdemo.vm:9000/
- Access key**: [Redacted]
- Secret key**: [Redacted]
- Bucket name**: [Redacted]



VMware Private AI Service 기능 #1

도메인 전문 지식 베이스 (Knowledge Base) 구성

- 지식 베이스 구축 – 임베딩 모델 및 Chunk Size 등 파라미터 설정 튜닝 기반 인덱싱

The screenshot displays the VMware Private AI Services web interface for configuring a Knowledge Base. The left sidebar shows the navigation menu with options like Data Indexing and Retrieval, Knowledge Bases, Data Sources, MCP, Agent Builder, and Observability. The main content area is titled 'KnowledgeBase' and includes a 'DELETE KNOWLEDGE BASE' button. It is divided into three panels: 'Knowledge Base Details', 'Link Data Sources', and 'Index settings' (highlighted with a red box). The 'Index settings' panel shows the index status as 'Completed' and lists the embedding model, chunk size, chunk overlap, and text splitter. Below these panels is a 'Document List' table showing indexed documents.

Document name	Status	File type	Data source	Processed at
A	Indexed	PDF	KNOC Sample Data	Apr 2, 2026, 5:28:18 PM
C	Indexed	HTML	KNOC Sample Data	Apr 2, 2026, 5:28:19 PM
E	Indexed	PDF	KNOC Sample Data	Apr 2, 2026, 5:28:20 PM



Agent Builder

- Model Endpoint, Instruction, Tools(Knowledge Base 등)를 결합한 맞춤형 AI 에이전트 생성

The screenshot displays the 'Create Agent' interface in the VMware Private AI Services console. The interface is divided into two main panels. The left panel shows a sidebar with navigation options: Data Indexing and Retrieval, MCP, Agent Builder, and Observability. The right panel is the 'Create Agent' form, which has three tabs: Agent Details, Tools, and View Configuration Code. The 'Agent Details' tab is currently selected, showing fields for Name, Description, Model endpoint, Instructions, and Delivery method. The 'Tools' tab is also visible, showing a list of Knowledge Bases with configuration options like Chunks to retrieve, Similarity cutoff, and Data references. A red box highlights the 'Model endpoint', 'Instructions', and 'Delivery method' fields in the 'Agent Details' tab. Another red box highlights the 'Knowledge Bases' section in the 'Tools' tab, which includes a table with columns for Name, Chunks to retrieve (Top K), Similarity cutoff, and Data references.

Name	Chunks to retrieve (Top K)	Similarity cutoff	Data references
Search Knowledge Base Index "KNOG-KnowledgeBase"	1	0.1	No reference



Agent Builder

- 목적 기반 에이전트 선별 - 업무 시나리오 및 사용자 의도에 최적화된 에이전트 선택

vmw Private AI Services

Logout

Namespace:

<<

Agent Builder > Agents

Data Indexing and Retrieval

Knowledge Bases

Data Sources

MCP

MCP Servers

Tools Gallery

Agent Builder

Agents

Observability

Metrics

Agents

CREATE AGENT

REFRESH AGENTS

	Agent name	Status	Agent ID	Model endpoint	Created on
	-Assist	Available	02ff01d5-00a9-4c29-b1cc-36c05c91a9bc	hugging-quants/meta-llama-3.1-70b-instruct-awq-int4	Apr 3, 2026, 11:27:34 AM
	-Safety	Available	12059ec9-3226-45ec-a6d6-5a8bd4da8277	hugging-quants/meta-llama-3.1-70b-instruct-awq-int4	Apr 3, 2026, 11:31:49 AM
	test	Available	e44cef80-ec8c-478e-8d91-bb1382f55a2d	hugging-quants/meta-llama-3.1-70b-instruct-awq-int4	Apr 3, 2026, 2:49:00 PM
	eng test	Available	0a08c149-65bb-4963-9353-f5b419483fbd	hugging-quants/meta-llama-3.1-70b-instruct-awq-int4	Apr 3, 2026, 3:08:02 PM

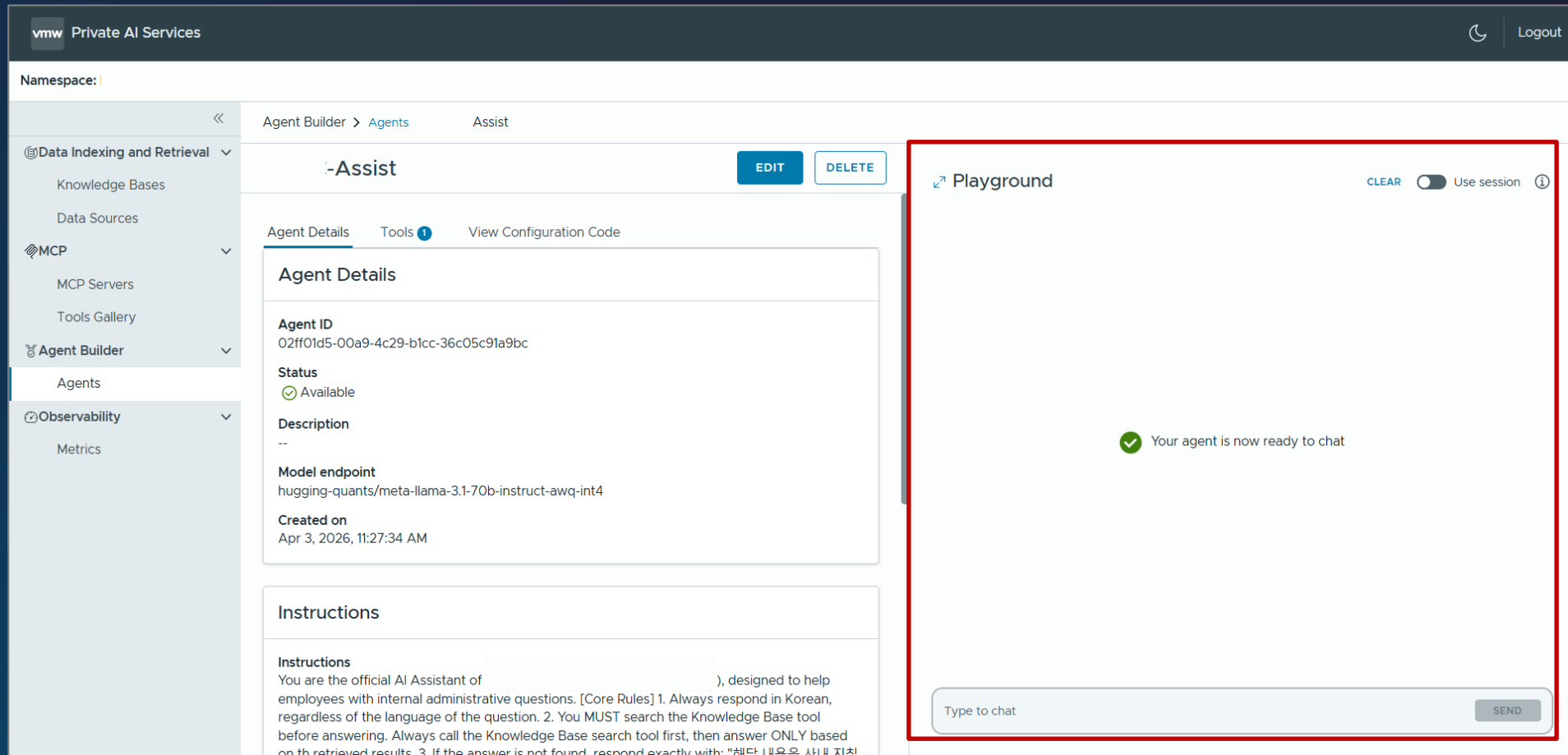
Manage Columns

Objects per page 10 4 objects



Agent Builder

- 대화형 인터페이스 상호 작용 – 자동 생성되는 AI Playground 환경에서의 실시간 질의응답 테스트



VMware Private AI Service

Demo 시연



AI Agent

- 구축된 지식 베이스 기반 실제 질문 입력
- 관련 문서 내용 인용, 정확한 답변 제공
- RAG 기반 전문 질의응답 시스템 시연

KEY POINT

- Knowledge Base 기반 전문 답변 생성
- 사내 규정 및 기술 문서를 실시간으로 참조
- Data Reference 기능으로 답변의 출처를 명시하여 신뢰성 확보

The screenshot displays the 'Playground' interface of the VMware Private AI Service. At the top right, there are controls for 'CLEAR', a toggle for 'Use session', and an information icon. The conversation area shows a user input (You) asking for information on security vulnerability scanning targets and scope. The AI assistant (Assist) responds with a detailed answer in Korean, listing various systems and components. Below the response, a 'HIDE DATA REFERENCE' link is visible. A 'Document 1' section provides specific metadata for the referenced document, including origin, retrieval ID, document ID, score, and media type.

Playground CLEAR Use session ⓘ

You
전사 보안 취약점 점검 대상 및 범위에 대해 알려줘

Assist
전사 보안 취약점 점검 대상 및 범위는 다음과 같습니다.

- [redacted]의 외부 서비스 웹 애플리케이션, 내부 비즈니스 시스템 웹 애플리케이션, 관련 네트워크 인프라, 서버 그룹
- 범위: 주요 웹 서비스의 [redacted] 취약성, 서버 OS 및 미들웨어 구성 취약성, 네트워크 보안 장치 우회 가능성에 대한 종합 점검

[HIDE DATA REFERENCE](#)

Document 1

- **Origin_name:** eng2.txt
- **Origin_ref:** test/eng2.txt
- **Retrieval_id:** ref01
- **Document_id:** e79e4eef-16df-4b49-bb88-9cc91dc0b128
- **Score:** 0.461
- **Media_type:** TXT



AI Agent

- 데이터 소스 연결 없이 LLM 자체 파라미터로 답변 생성
- 범용 정보로 답변, System Prompt 설정하여 API Endpoint 형태로 3rd Party App과 연계 가능

Agent Builder > Agents >

✧ Playground

CLEAR ☐ Use session ⓘ

안전대 설치 시, 작업자들의 안전을 확보하기 위해 반드시 작업자들의 의견을 수렴하고, 작업자들의 안전을 우선시합니다.

만약, 안전대 부착 설비 설치 기준 높이에 대한 더 자세한 정보가 필요하시면, 한국산업안전보건공단(KOSHA)의 가이드라인을 참조하시기를 권장합니다.

You

강풍 시 고소 작업 중지 기준 풍속은?

Safety

강풍 시 고소 작업 중지 기준 풍속은 일반적으로 15m/s 이상입니다.

다음은 강풍 시 고소 작업 중지 기준 풍속을 준수하기 위한 절차입니다.

- **고소 작업 전 풍속 확인:** 작업 시작 전 풍속을 확인하고, 15m/s 이상인 경우 작업을 중지합니다.
- **풍속 모니터링:** 작업 중 풍속을 지속적으로 모니터링하고, 15m/s 이상이 되면 즉시 작업을 중지합니다.
- **안전한 작업 환경 확보:** 작업 환경을 안전하게 확보하고, 작업자들이 안전하게 작업할 수 있는 조건을 확보합니다.
- **풍속에 따른 작업 중지 기준:** 풍속에 따른 작업 중지 기준을 명확하게 정의하고, 작업자들이 이를 준수할 수 있도록 교육합니다.

중요: 풍속이 15m/s 이상인 경우, 고소 작업을 중지하고 작업자들을 안전한 장소로 대피시켜야 합니다.

Type to chat

SEND

↗ Playground

CLEAR ☒ Use session ⓘ

You

황화수소(H2S) 노출 기준치와 응급 처치 방법은?

Safety

황화수소(H2S) 노출 기준치는 다음과 같습니다.

- 10 ppm 이하: 일반 작업 가능
- 10-20 ppm: 작업자에게 개인보호구 착용 필요
- 20-50 ppm: 작업자에게 개인보호구 착용 필요, 작업자 수 제한
- 50 ppm 이상: 작업 중단, 작업자 즉시 대피 필요

황화수소(H2S) 중독 발생 시 응급 처치 방법

- 환자에게 신선한 공기를 제공하고 즉시 대피시킴
- 환자가 의식이 있는 경우, 즉시 의료진에게 연락하고 의료 처치를 받음
- 환자가 의식이 없는 경우, 즉시 심폐소생술(CPR)을 실시하고 의료진에게 연락함
- 환자가 호흡 곤란을 겪는 경우, 즉시 산소.mask를 착용시킴
- 환자가 구토를 하는 경우, 즉시 구토물을 제거하고 의료진에게 연락함

중요한 안전 주의사항

- 황화수소(H2S) 작업 시, 작업자들은 반드시 개인보호구를 착용해야 함
- 황화수소(H2S) 작업 시, 작업자들은 반드시 작업장의 환기 상태를 확인해야 함
- 황화수소(H2S) 작업 시, 작업자들은 반드시 응급 처치 장비를 준비해야 함



AI Agent – MCP

- MCP 설정 기반의 사전 설정

vmw Private AI Services

Namespace: pais-datasolution-bf7n2

Agent Builder > Agents > Mixtral-KO

Left sidebar menu:

- Data Indexing and Retrieval
 - Knowledge Bases
 - Data Sources
- MCP
 - MCP Servers
 - Tools Gallery
- Agent Builder
 - Agents
- Observability
 - Metrics

Dialog Title: Add Tools

Select approved tools from MCP servers or knowledge bases from Data Indexing and Retrieval.

1 selected Show only selected

☐	Tool name	Description	MCP Server
☒	Search Knowledge Base Index "KB-Demo"	Search for relevant content in the knowledge base index linked to this tool. The Knowledge Base Index linked to this tool is called "KB-Demo".	Knowledge Bases
☐	Search Knowledge Base Index "KB-Demo-Korean"	Search for relevant content in the knowledge base index linked to this tool. The Knowledge Base Index linked to this tool is called "KB-Demo-Korean".	Knowledge Bases
☐	Search Knowledge Base Index "AVI-KO"	Search for relevant content in the knowledge base index linked to this tool. The Knowledge Base Index linked to this tool is called "AVI-KO".	Knowledge Bases

☒ 1 Manage Columns Objects per page: 10 1 - 3 of 3 objects

CANCEL ADD

Background chat area: Save the agent first to chat with the agent SEND

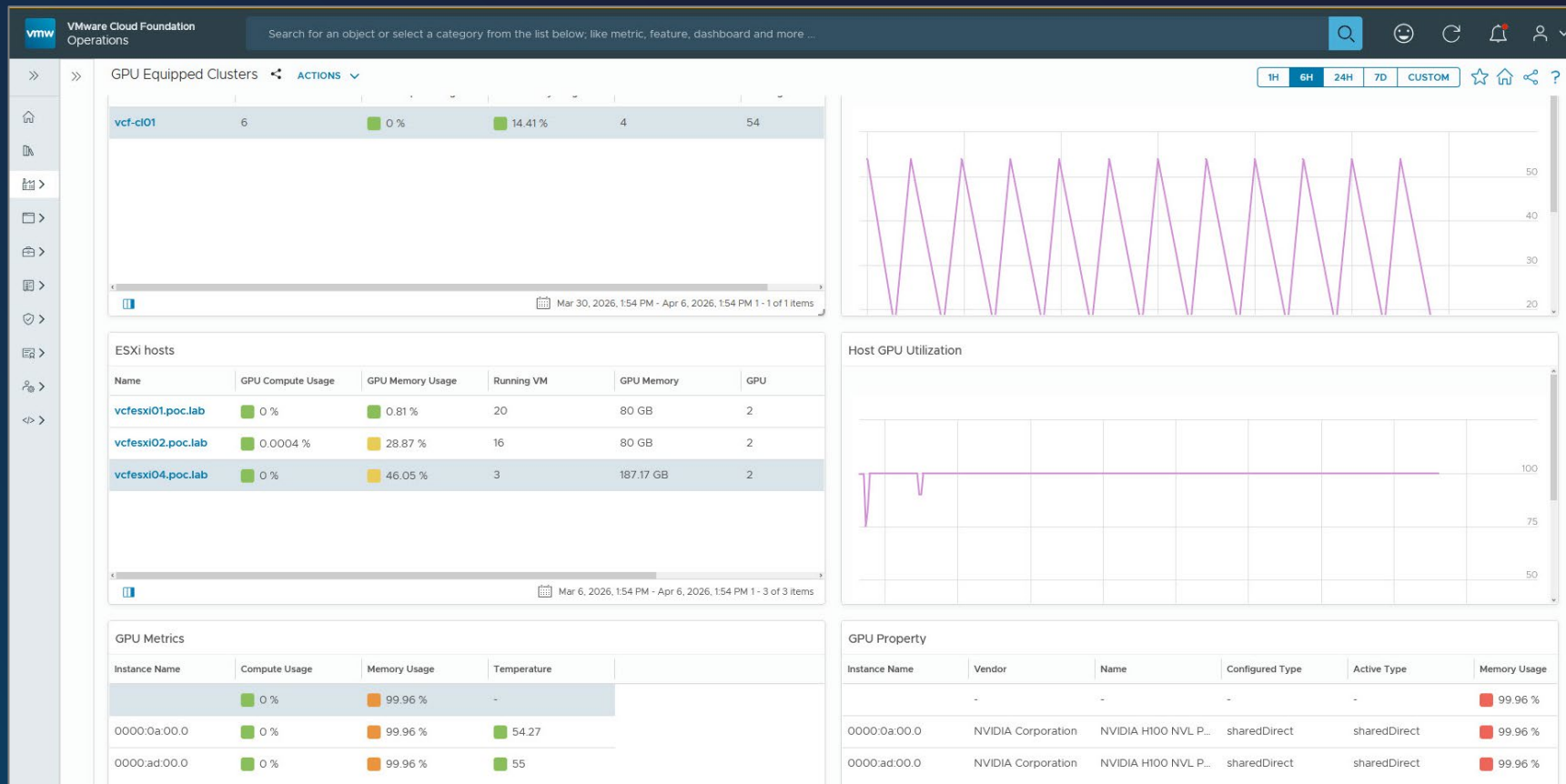


AI Agent - MCP

VMware Private AI Service 기능 #3

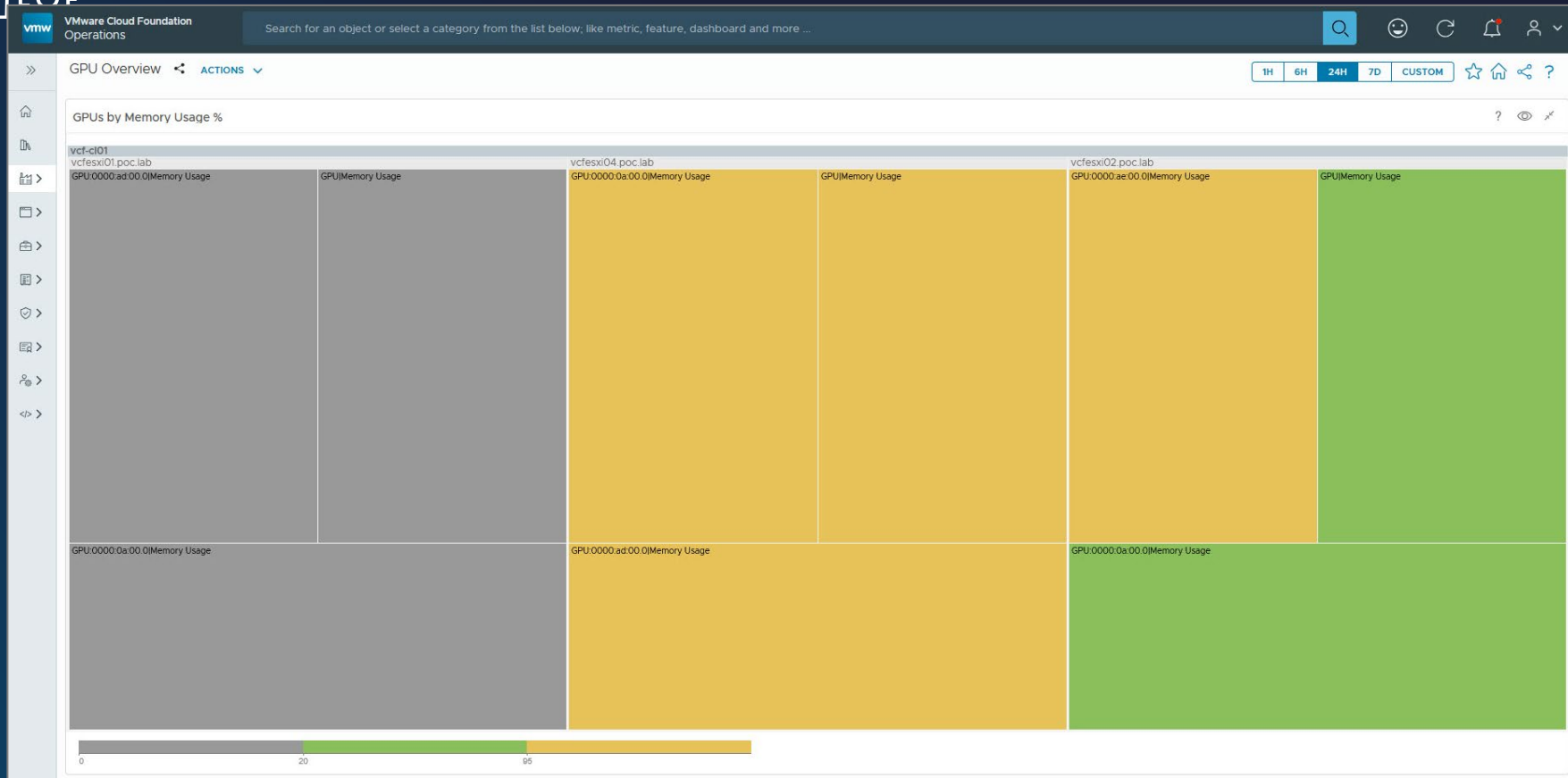
AIOps 및 모니터링

- VCF Operations를 통한 GPU 상태 모니터링 → 안정적인 AI 운영 (AIOps 지원)
- 연산 활용도/메모리 사용량/온도 대시보드 제공



AI Ops 및 모니터링

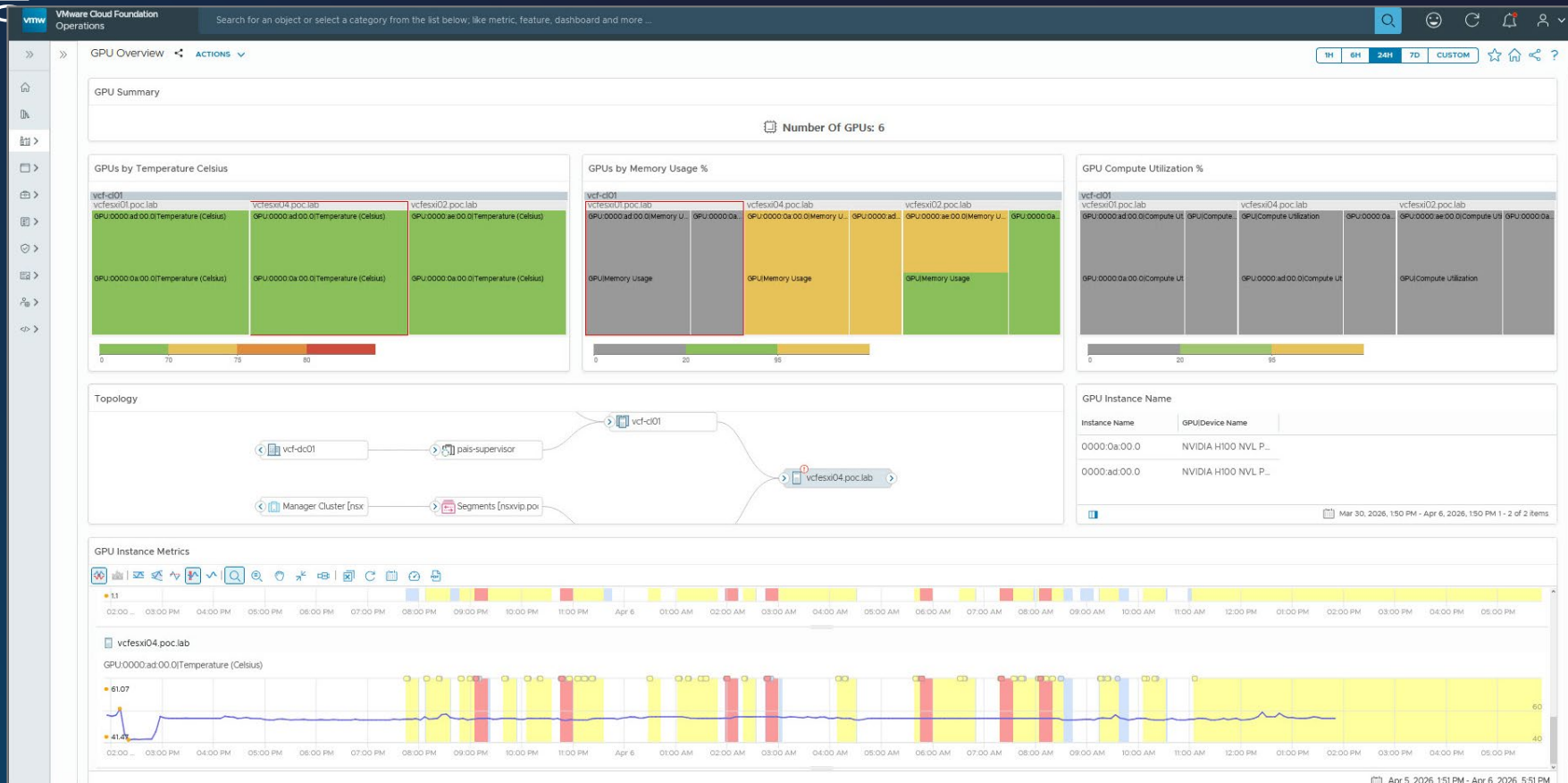
- Heat-Map 기반 GPU 메모리 점유율 분석
- 전체 GPU 자원의 메모리 사용 현황을 컬러맵으로 시각화하여 자원 편중 현상 및 가용 상태를 한눈에 파악



VMware Private AI Service 기능 #3

AI Ops 및 모니터링

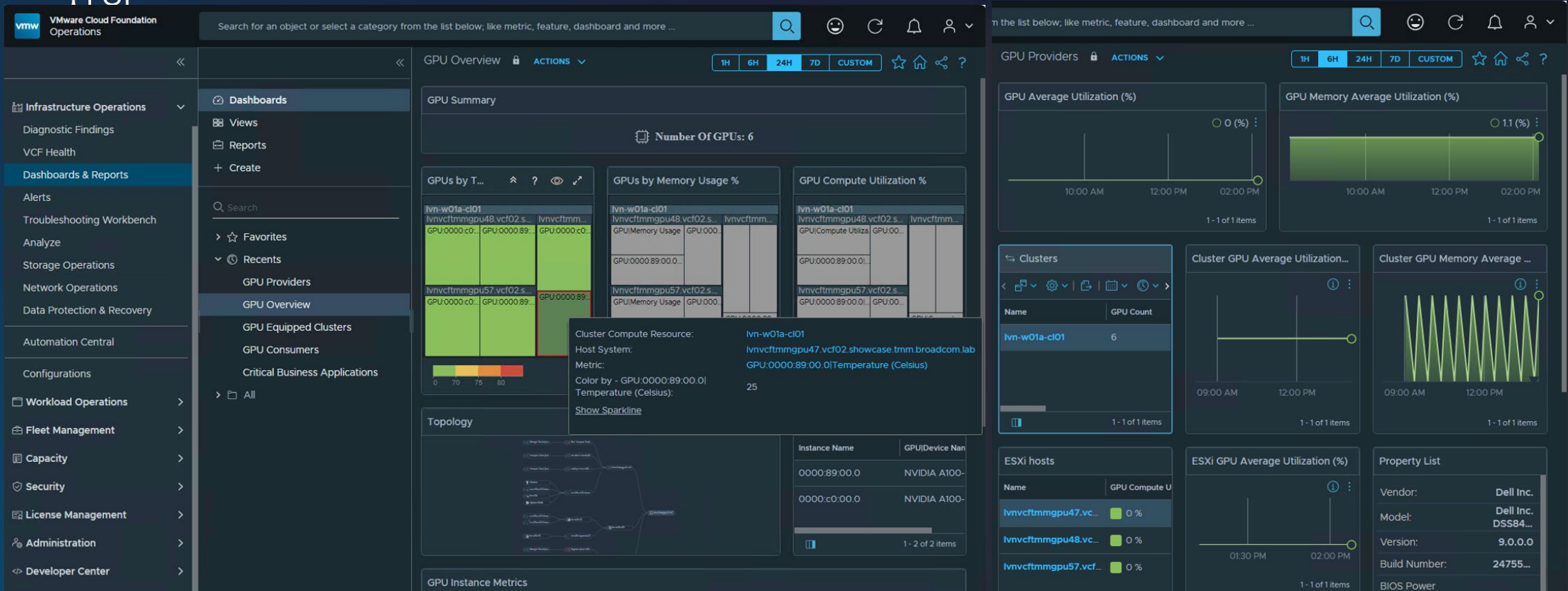
- GPU 인프라 토폴로지 및 상태 요약 대시보드
- GPU 온도, 메모리, 연산지표의 추이를 통합 관리 / 물리적 위치와 논리적 연결 구조를 토폴로지 뷰로 제공



VMware Private AI Service 기능 #3

AIOps 및 모니터링

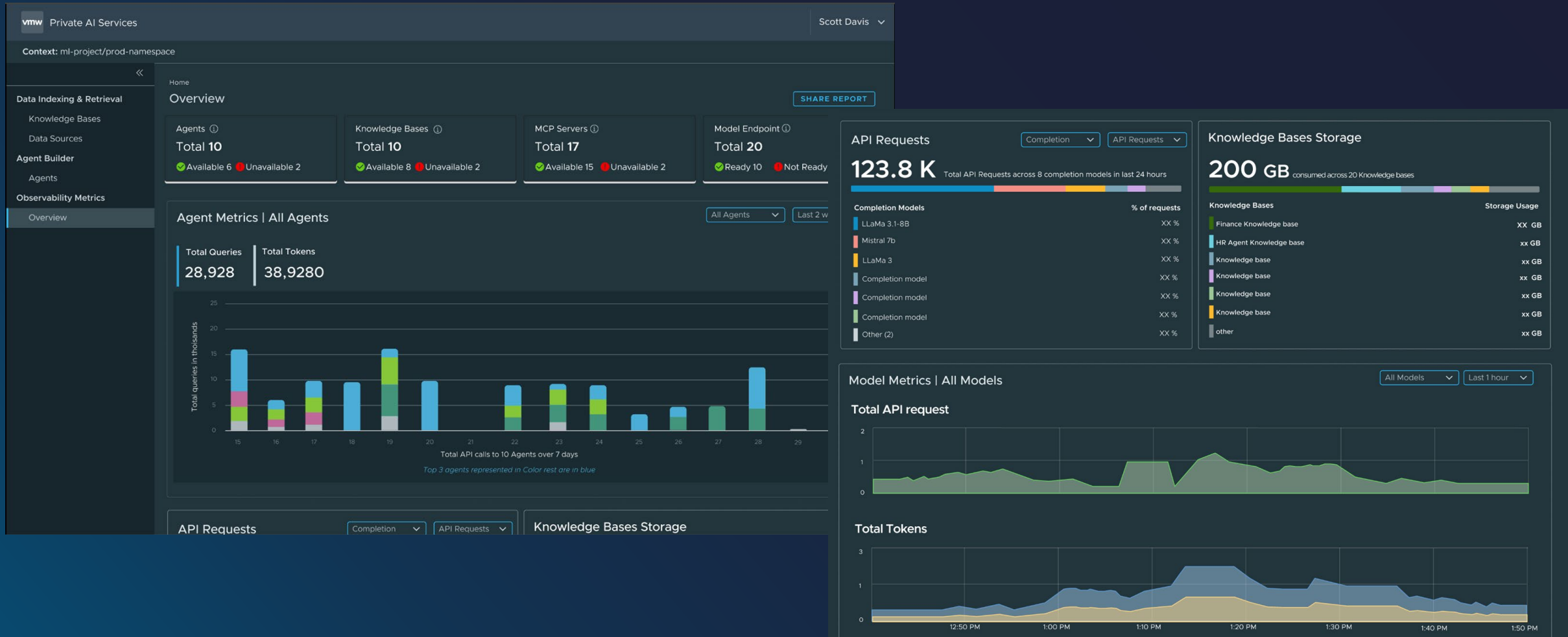
- GPU 공급자별 성능 트렌드 분석
- 일정 기간 동안의 GPU 평균 활용도 및 메모리 사용량 추이 분석, 인프라 증설 및 최적화 의사결정 지원



VMware Private AI Service 기능 #3

AI Ops 및 모니터링

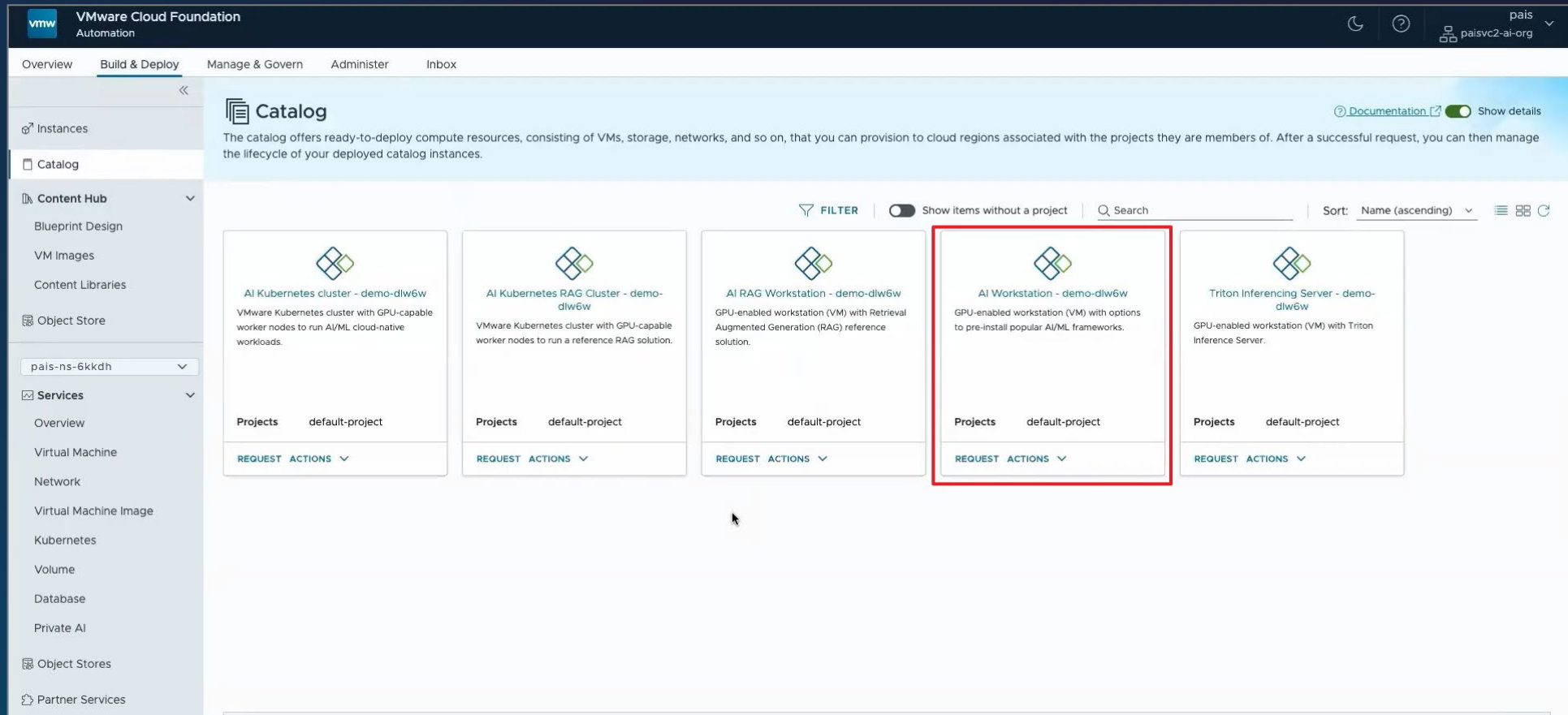
- 자체 메트릭 기반 Total Queries / Total Tokens / API Request 등 가시성 제공



VMware Private AI Service 기능 #4

ML/DL 환경 배포

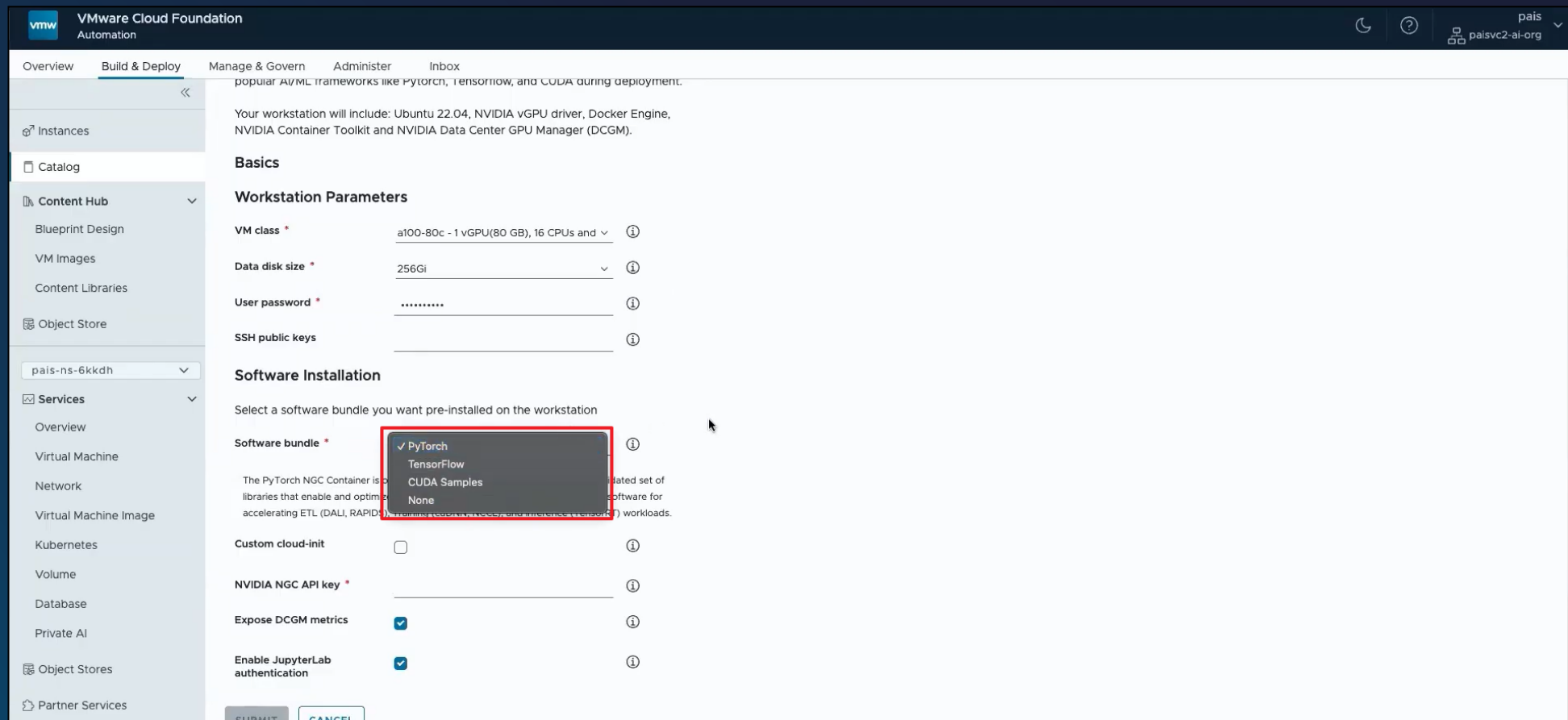
- 검증된 AI 워크스테이션 및 TKG 클러스터 템플릿을 카탈로그 형태로 제공
- 복잡한 인프라 설정 없이 목적에 맞는 AI 환경을 즉시 선택하여 요청 가능



VMware Private AI Service 기능 #4

ML/DL 환경 배포

- 사용자 요구사항에 최적화된 GPU VM Class(vGPU 프로파일)와 가용 디스크 용량을 직접 지정
- PyTorch, TensorFlow등 핵심 ML/DL 프레임워크를 선택하여 필요한 소프트웨어 스택 자동 설치



VMware Private AI Service 기능 #4

ML/DL 환경 배포

- 요청된 AI 워크스테이션이 정상적으로 배포되었으며, GPU 가속 최적화 상태임을 확인
- 배포 상세 내역을 통해 할당된 리소스 정보와 JupyterLab 등 즉시 실행 가능한 어플리케이션 접속 링크

The screenshot displays the VMware Cloud Foundation Automation interface. The top navigation bar includes 'Overview', 'Build & Deploy', 'Manage & Govern', 'Administer', and 'Inbox'. The user 'palsvc2-ai-org' is logged in. The main content area shows the deployment details for 'demo-dlvm', which has a 'Create Successful' status. A 'Show details' toggle is visible. The deployment information includes:

Project	Expires on	Owner
default-project	Never	pals

Additional details include:

- Blueprint:** [AI Workstation - demo-dlw6w_version:1](#)
- Created on:** Dec 17, 2025, 1:59:55 PM
- Requestor:** pals
- Catalog item:** AI Workstation - demo-dlw6w

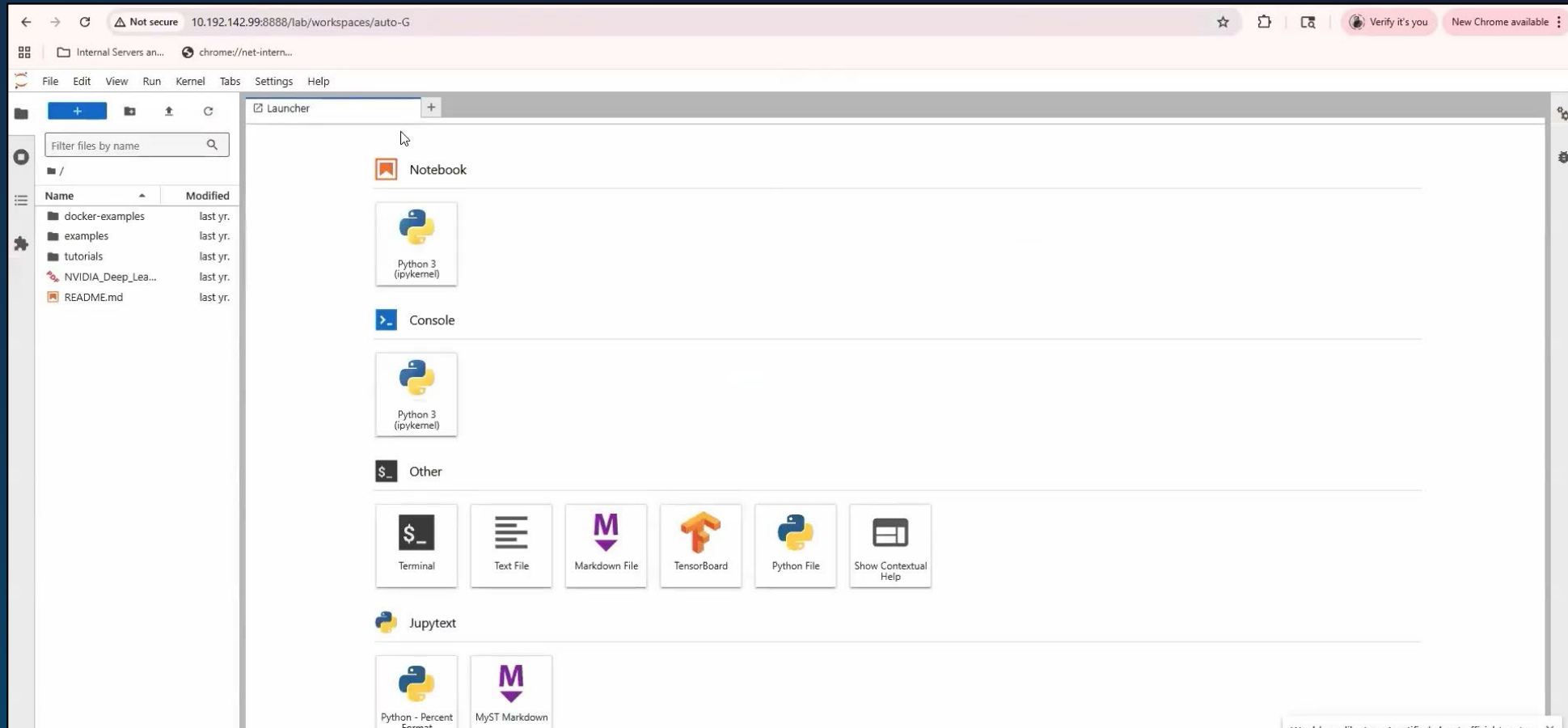
The 'Overview' tab is selected, showing the 'AI workstation' section. It includes a 'PyTorch' section with a description of the NGC container and a link to the NVIDIA catalog. Below this is an 'Applications' section listing 'JupyterLab' as an available application. A 'CLOSE' button is located at the bottom left of the console window.



VMware Private AI Service 기능 #4

ML/DL 환경 배포

- 배포 후, 별도의 추가 설정 없이 웹 브라우저를 통해 JupyterLab 환경에서 개발 시작
- 터미널, 노트북 등 다양한 도구와 사전에 선택한 프레임워크가 완벽히 구성된 최적의 환경 제공



VMware Private AI Service 기능 검증 기대효과 및 결론

70%

데이터 분석 시간 단축

2x

GPU 자원 활용 효율 향상

100%

보안 컴플라이언스 기준 준수

PoC 핵심 검증 항목

- ✓ VCF 9.0 배포 자동화 및 SDDC 통합 관리 기능 검증
- ✓ Private AI 기반 vLLM 서비스 안정적 운영
- ✓ GPU 가속 VKS 클러스터 및 MLOps 자동화 파이프라인
- ✓ NSX 마이크로세그멘테이션 기반 Zero-Trust 보안 검증
- ✓ vSAN 고성능 스토리지 AI 데이터 처리 성능 검증

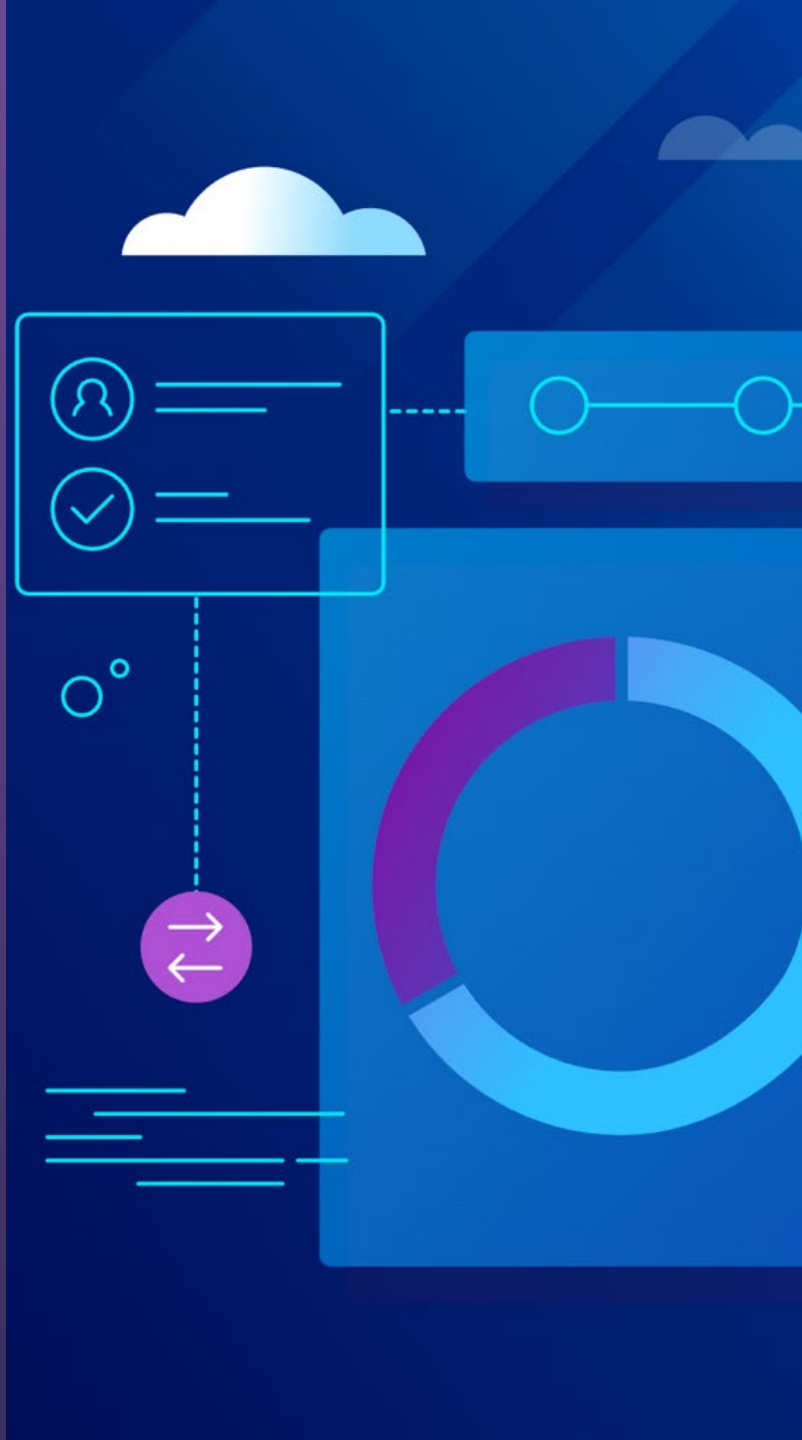
PoC 완료 후 Next Steps

- ✓ PoC 결과 보고서 작성 및 최종 시연 발표
- ✓ VCF 9.0 본 도입 아키텍처 상세 설계 착수
- ✓ Broadcom 라이선스 최적화 협의 (통합 유지보수 연계)
- ✓ 하드웨어 교체 계획 수립 (vSphere 8.x EOS 대비)
- ✓ 파일럿 운영 환경 구성 및 단계적 확산 계획 수립





Why Private AI on VCF?



Why Private AI on VCF?

VCF Private AI Service는 오픈소스 조합의 복잡성과 불확실성을 제거하고, 엔터프라이즈급 RAG AI 환경을 빠르고 안정적으로 제공합니다

Values



구축 기간 단축

핵심 구성 요소
즉시 사용 가능



검증된 안정성

VCF 플랫폼 내 검증된
구성 요소 제공



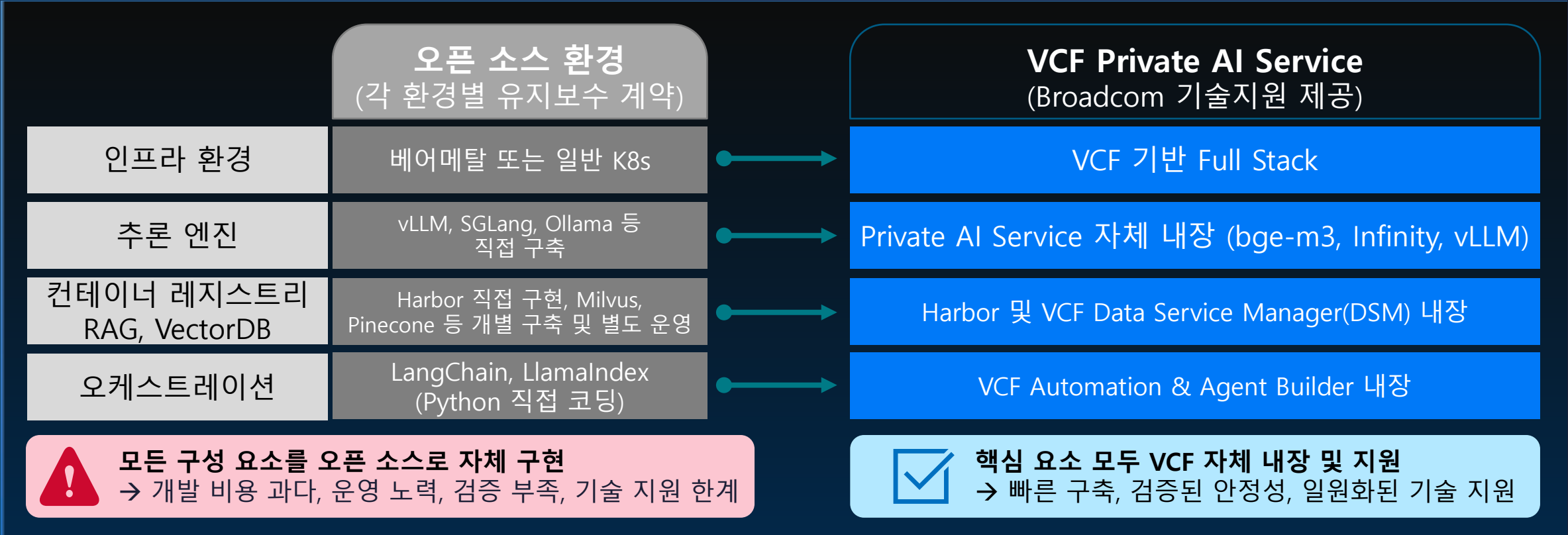
운영 효율성 향상

운영 부담 최소화,
데이터, 서비스에 집중



일원화된 기술 지원

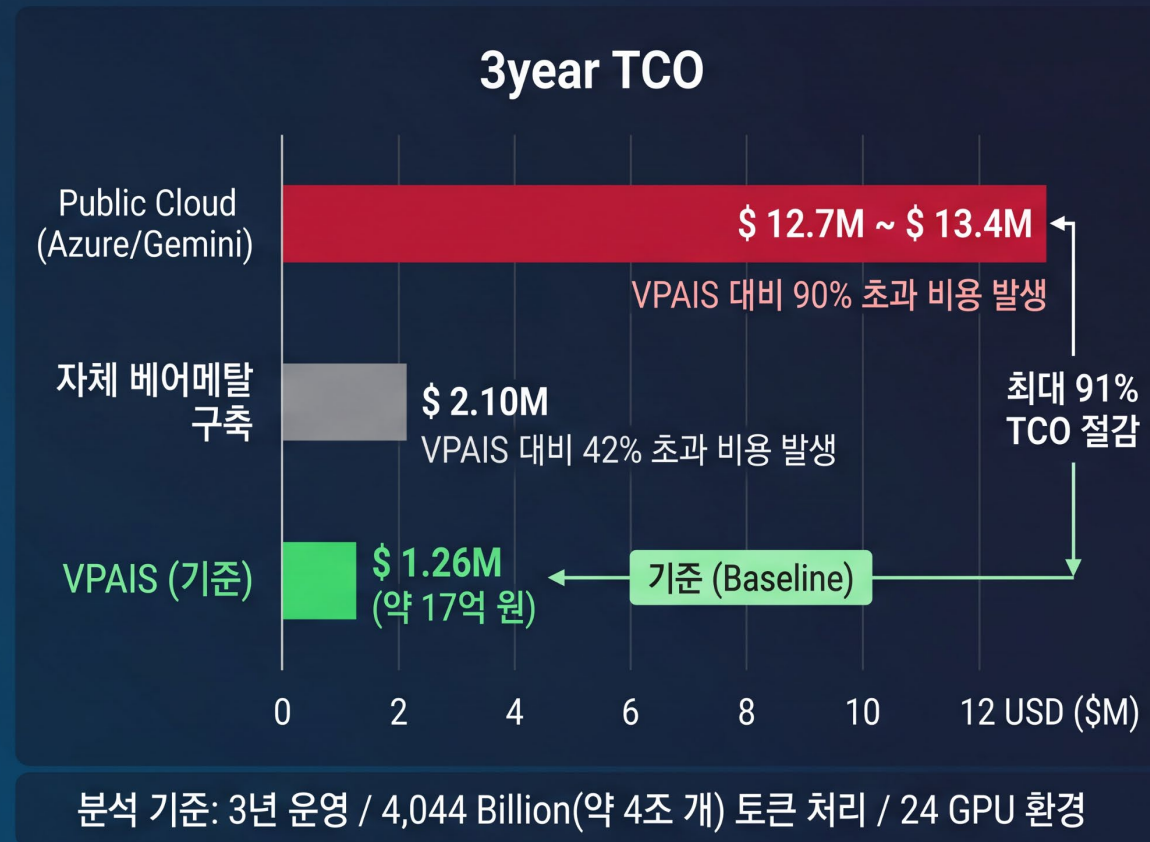
Broadcom 에서 통합
기술 지원 제공



TCO 91% 절감의 비밀: 변동성(OPEX)에서 통제 가능한 자산(CAPEX)으로

동일한 4조 개의 토큰, 전혀 다른 청구서 - 인프라 '구조'가 비용을 결정합니다.

Objective Data Chart (Proved by Numbers)



2 Structural Root Causes of Cost Gap

① 퍼블릭 클라우드의 함정 (OPEX Risk)



원인: **API 호출 토큰**에 비례한 **기하급수적 종량제 과금** 및 막대한 S/W 라이선스 비용

결과: 자율형 에이전트 시대에 **예측 불가능한 IT 예산 붕괴** 초래

② 베어메탈(물리 서버)의 비효율 (Resource Silo)



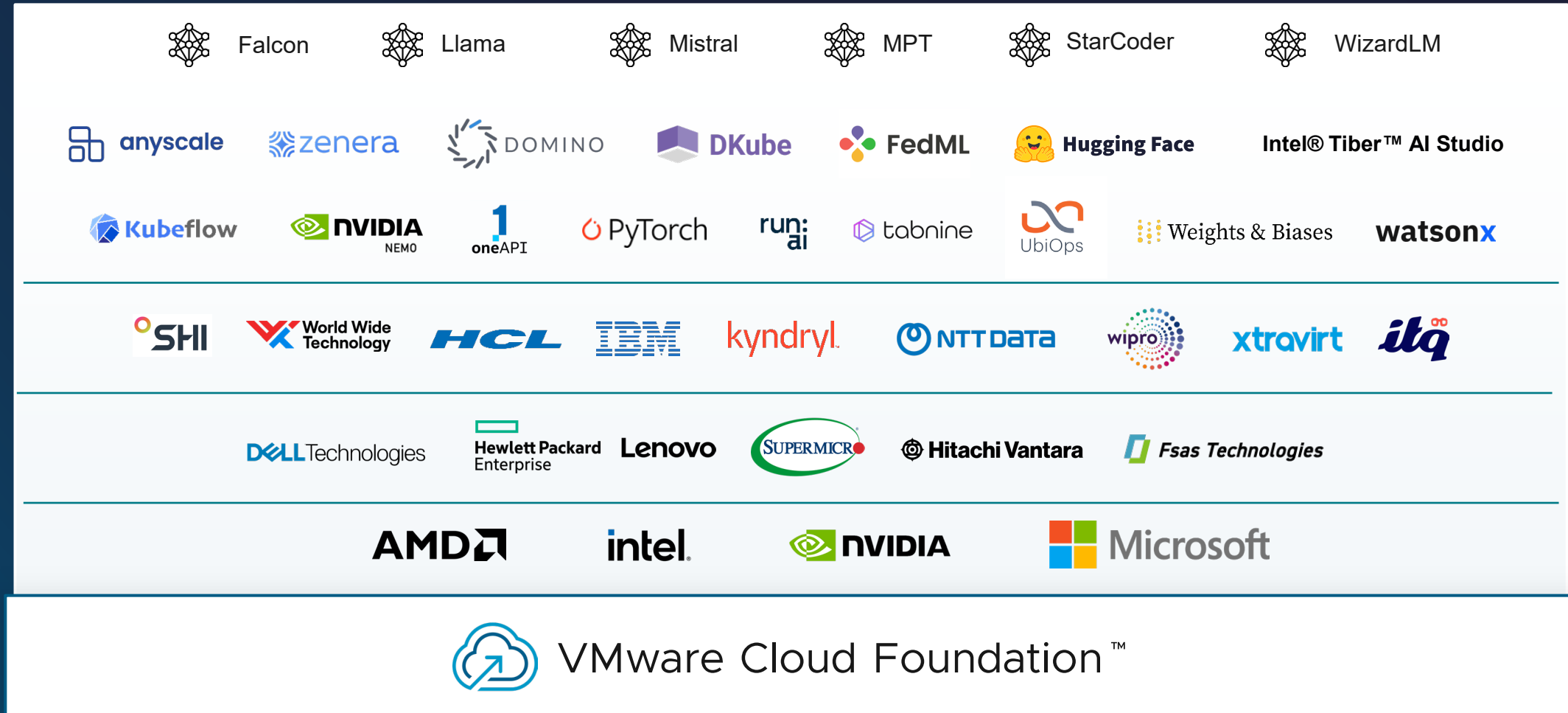
원인: 부서별 **GPU 파편화 할당**으로 인한 유휴 자원 발생 및 하드웨어 중복 투자

결과: **VCF 9**의 가상화 기반 **GPU 풀링(Pooling)** 기술 적용 시, 순수 하드웨어 요구량을 베어메탈의 **절반 이하로 감축** 가능

출처: Broadcom TCO Calculator



VMware Private AI Open Ecosystem



Key Takeaway & Call to Action

온프레미스 AI 인프라를 단일 플랫폼으로 통합하는 VCF Private AI Service를 바로 배포해보세요!

1

망분리 규제 준수 및 AI 거버넌스 확보

프라이빗 클라우드 환경에서 데이터 분석 및 AI Agent 고도화

2

AI 워크로드까지 확장되는 고가용성

vGPU에도 vMotion / HA / DRS 등 고가용성 기능 완벽 지원

3

수 분 내 AI 서비스 배포 및 운영

개발/운영 서비스, 추론 엔진까지 카탈로그 선택 한 번으로 빠른 배포



